Predictors of Earnings Risk with Machine Learning

Ethan Ballou[*] Scott Drewianka[*]

January 26, 2026

Abstract

This paper looks at the determinants of lifetime earnings risk under a Restricted Income Profile (RIP) model using traditional and machine learning methods such as lasso and SHAP values. The paper builds on the work of Drewianka and Oberg (2025) which uses a moment condition approach derive a parameter that captures permanent income risk. The paper finds that education and age are important in explaining lifetime earnings risk. The paper also finds that macroeconomic variables such as probability of recession and real GDP growth are important and along with state controls may further imply a role of government policy. Finally, the paper finds that occupation controls are important while industry controls do not appear to play a strong role.

Keywords: machine learning, restricted income profile, earnings instability, risk
JEL Codes: D8, J0, D3

To Do

1. Rewrite Introduction
2. Edit Theoretical Framework
3. Rewrite Data
4. Rewrite Empirical Strategy
5. Rewrite Results
6. Rewrite Conclusion
7. Consistent formatting of bibliography

- Use old sections as well as suggested edits as an outline for rewritten sections

[*] University of Wisconsin - Milwaukee

1 Introduction

Understanding the nature of earnings risk is central to both individual decision making and policy. People are often thought to be risk averse not risk neutral and therefore earnings risk has an effect on how individuals make intertemporal decisions such as saving. Especially on the analysis side, structural life cycle models have to make assumptions about the nature of earnings dispersion and risk. Misspecifying the earnings process can lead to inaccurate or incorrect conclusions in many cases and this is why understanding the nature of earnings risk and the earnings process is so important.

To better pin down the quantitative importance of shocks versus profile heterogeneity, this paper builds on the work of (Drewianka and Oberg 2025, 2019) [5, 6] which uses a moment condition approach to test for heterogeneity in expected income processes. The paper uses a Restricted Income Profile (RIP) model to estimate lifetime earnings risk and then tests for heterogeneity in expected income processes using a moment condition approach.

Restricted Income Profile (RIP) models posit that all individuals with the same observed characteristics follow the same earnings trajectory. Deviations then being attributed to shocks to the individuals income process. In RIP, dispersion is then attributed to the variance of these shocks. (Abowd and Card, 1989; Meghir and Pistaferri, 2004) [1, 13] This means that the variance of shocks and variance in observed characteristics are sources of dispersion in earnings. Some of the variance is due to shocks the rest of the variance is due to the different expected income processes. (Guvenen 2007; Baker 1997) [7, 2]

This paper tests for predictors of lifetime earnings risk as parameterized by the model in Drewianka and Oberg (2025) [5]. The paper hopes to find which characteristics are important in explaining the variance of shocks in the RIP framework.

This paper finds that education and age are important in explaining lifetime earnings risk. The paper also finds that macroeconomic variables such as probability of recession and real GDP growth are important and play at least a small role and that the inclusion of state controls may further the hypothesis of government policy playing a role. Finally, the paper finds that occupation controls are important while industry controls do not appear to play a large role.

This paper contributes to multiple strands of literature. The first is the literature around (RIP) models and their use in life cycle models. (MaCurdy 1982; Abowd and Card 1989; Hryshko 2012). [12, 1, 10] In addition to that it also contributes to a second strand of literature focusing on lifetime earnings risk. (Drewianka and Oberg 2025; Guvenen 2007; Meghir and Pistaferri 2004). [5, 8, 13] Finally the paper contributes to the literature on machine learning and its use in economics by utilizing neural networks and SHAP values to interpret the results (Lundberg and Lee, 2017). [11]

The rest of the paper is structured as follows. Section 2 discusses the model and data, Section 3 discusses the empirical strategy, Section 4 discusses the results, and Section 5 concludes the paper.

1. preview the results at the end
2. Also preview the methods used a little more
3. spend a little more time explaining the contribution of this paper specifically, in both using some fancier methods for lack of a better term to identify cove variance with this risk, but also the importance of being able to characterize people who are at an elevated risk in both permanent and transitory horizons

2 Theoretical Framework

As mentioned before, the model used is the same model from Drewianka and Oberg (2025) [5]. The model is a standard RIP model where people's income are modeled as a function of characteristics, but where there are no individual level effects. Then each individual has a deviation from the expected income process which is modeled as a shock. For example all 22 year old men from the US will have the same expected income process based off of those characteristics. However their actual income will deviate from this expected income process due to shocks.

$$u_{it} = \pi_{it} + \nu_{it} \tag{1}$$

$$\pi_{it} = \rho\, \pi_{i(t-1)} + \eta_{it}, \tag{2}$$

u_{it} is the shock or deviation from the expected income process for a given person i in time t. This process of shocks can be decomposed into two types of shocks: persistent and temporary shocks as outlined in equation 1. π_{it} is the cumulative effect from the persistent shocks which are further modeled in equation 2 while ν_{it} are the temporary shocks. The temporary shocks along with η_{it} are assumed to be mean zero and independent across individuals and time. The persistent shocks are further modeled as an autoregressive process with ρ being the persistence of the shocks and η_{it} being the actual shock.

To get a measure of earnings risk the variance of η_{it} across time is calculated. This is capturing the correlation of permanent income shocks across time for each individual. To calculate this the following equation is used:

$$\Omega_i(t,t+k) \equiv u_{i(t+k)} - u_{it} \tag{3}$$

$$\Omega_i(t,t+k) = \nu_{i(t+k)} - \nu_{it} + \sum_{j=1}^{k} \eta_{i(t+j)}. \tag{4}$$

This omega equation differences u_{it} across a window t to t+k. This is then multiplied by another omega function across the window t-j to t+k+q. This will result in the persistent shocks in the interval (t, t+k) showing up twice while the other persistent shocks don't necessarily show up twice. This results in the variance of the shocks across time k as follows:

$$E\Big[\tfrac{1}{k}\,\Omega_i(t,t+k)\,\Omega_i(t-j,t+k+q)\Big] = \sigma^2_{\eta_i}, \tag{5}$$

$$\gamma_{itjq} \equiv \tfrac{1}{2}\,\Omega_i\big(t,\, t+2\big)\,\Omega_i\big(t-j,\, t+2+q\big). \tag{6}$$

This is then further used to get the following moment condition for the variance of η_{it} that is a function of (j+k+q) and j and can be estimated in a regression framework to get a measure of lifetime earnings risk γ_{itjq}. For more on the derivation see Drewianka and Oberg (2025) [5]:

$$E[\gamma_{itjq}] \approx \big[\delta^2\, E\pi^2_{i(t-j)}\,(j+k+q)\big] - \big[\delta\, s_i(t-j+1,t)\, j\big] + \big[\sigma^2_{\eta_i(t+1)} - \delta\, \sigma^2_{\eta_i(t+1)}\,(k+q)\big] \tag{7}$$

$$E[\gamma_{itjq}] = \sigma^2_{\eta_i(t+1)} + \big[\delta^2\, E\pi^2_{i(t-j)} - \delta\, \sigma^2_{\eta_i(t+1)}\big]\,(j+k+q) + \delta\,\big[\sigma^2_{\eta_i(t+1)} - s_i(t-j+1,t)\big]\, j \tag{8}$$

3 Data

The data used is the 1970-2023 waves of the Panel Study of Income Dynamics (PSID 2023). [14] The sample is of men between the ages of 22 and 69 with students being excluded. Many of the men in the sample come from the Survey of Economic Opportunity which has been shown to be representative of the main PSID sample (SEO; Hill 1992; Drewianka 2010). [9, 4] Income is calculated in 2015 dollars using the Consumer Price Index (CPI 2017). [3] The analysis uses several key variables, including education (categorized as less than high school, high school or some college, and bachelor's degree but less than a master's degree), age (with higher-order polynomial terms Age^2 and Age^3 to capture non-linear effects), and macroeconomic indicators such as the probability of a recession (P(Recession)) and real GDP growth. Individual-specific variables include fixed effects for wages, a moving average of income over the last five years (MA(Last 5 years income)), employment status (Employed), and job tenure (years in the current job). The primary outcome variables, gamma (γ)

and alpha (α), measure lifetime earnings risk. The gamma statistic measures the variance in permanent income shocks while the alpha statistic measures the variance in temporary income shocks. Both these statistics are calculated using the moment condition approach outlined in Drewianka and Oberg (2025). [5]. The income variable used is the log of the individual's hourly wage. The hourly wage is calculated by dividing annual labor income by the product of weeks worked and usual hours worked per week.

Additionally, the analysis incorporates fixed effects for occupation, industry, state of residence, and other controls such as year, race, and cohort. With the theoretical framework and data laid out, the following section turns to the empirical implementation of the model.

1. Mention CNEF and how it was combined with the PSID
2. Mention that macro economic variables are from another source and cite it
3. Discuss inconsistencies with some of the variables across years, general PSID BS, tenure, occ, ind, race even
4. Add a table with summary stats maybe? Describing people with certain combinations of characteristics, possibly a concern for overfitting with ML methods

4 Empirical Strategy

The empirical strategy of this paper is broken into two parts. The first part is estimating gamma and alpha, then contruct a weighted gamma and alpha for each individual in each year since both statistics are across i, t, j, and q. (i - individual, t - time, j and q - period or "window" used in calculation). The second part of the empirical strategy is estimating the then weighted values of each statistic using various models and variables to look for trends in the characteristics of individuals in regards to lifetime earnings risk.

For the estimation of both gamma and alpha, consolidation across j and q is done by using a mixed regression for the gamma and then a fixed effects regression to get the composite or weighted value for each person year.

The following mixed regression is used to estimate gamma based on the above equation. This is done to get an estimate of gamma for each individual in each year across j and q. A mixed regression used instead of standard OLS because $\sigma^2_{\eta_i}$ is correlated with both the constant and the coefficient on (j+k+q) as seen in equation 8 and therefore a mixed regression is used to account for this by allowing the residuals to be separately correlated with (j+k+q) for each individual i in year t. (For more on this see Drewianka and Oberg (2025) [5]) The mixed regression is as follows:

$$\gamma_{itjq} = \beta_0 + \beta_1 (j + k + q) + \beta_2 j + \epsilon_{itjq}, \tag{9}$$

$$\epsilon_{itjq} = \beta_{0it} + \beta_{1it} (j + k + q) + e_{itjq}. \tag{10}$$

The random effects component (j+k+q) models how each person-year combination may have a unique relationship with the (j+k+q) variable while the cov(unstructured) option in Stata allows for unrestricted correlation between the random intercept and slope. The standard errors are clustered at the person level to account for within-person correlation across time.

From here a regression is run across fixed effects for each combination of j and q with year effects absorbed. Weights are then calculated for each combination of j and q based on their accuracy in predicting gamma using inverse MSE. Specifically, after estimating the fixed effects model, the mean squared residual (MSE) is calculated for each (j,q) combination, then use the square root of the inverse of these values ($\sqrt{(1/MSE)}$) as weights. Then the weighted gamma is calculated consolidating across j and q so that there are only individual-year gamma values. Having obtained individual-year estimates of earnings risk, we now examine the determinants of this risk using several regression techniques.

The second part of the empirical strategy uses various models to estimate lifetime earnings risk (gamma). We begin with the simplest specification to establish a baseline before introducing more sophisticated selection and machine-learning methods. The first model is a standard OLS regression of the lifetime earnings risk variable, gamma, on various controls and variables. The second model is a stepwise regression which selects variables based on a p-value threshold. The third model is a lasso regression which penalizes the size of the coefficients to select variables. Finally, the fourth model uses a multi-layer perceptron and SHAP values to interpret the results.

The OLS regressions are fairly standard however the stepwise and lasso models do have components that are worth mentioning. For the stepwise regression model a cutoff p-value of 0.05 is used to select variables. The model removes the least significant variable (or group of variables in the case of a set of controls) in rounds until it reaches the cutoff. 0.05 was selected so that the model would still select some variables but the rankings of the many of the variables would be clear. A higher cutoff and most of the variables would be selected and cardinal rankings would not be visible. Too low of a cutoff value and many of the variables would not be selected at all.

As for the lasso model, the model is set up to select variables based on the Bayesian Information Criterion (BIC) which is a common method for selecting variables in penalized regression. While the model selects lambda based on cross-validation, the selected model isn't very relevant to the analysis as the rankings are so the selected model isn't discussed. It is worth mentioning however that the lasso model is a penalized regression and therefore some weight is assigned to the size of the coefficients when selecting variables. This is different from the stepwise regression which is indiscriminate regarding the size of the coefficient and only considered the significance of a variable.

Finally the multi-layer perceptron model is a neural network that is used to estimate lifetime earnings risk. The model is trained on the same variables and data used by the OLS, stepwise, and lasso models. The model is 4 dense layers (excluding the input layer) with 1000 nodes each and a linear output layer at the end. The model used a sigmoid activation function in the hidden layers to allow for continuous support. This was done instead of a ReLU activation function to allow for more definition in the parameters instead of some parts of the perceptron being "dead" and not contributing and complicating the SHAP value interpretation. This model size and structure was selected as it achieved the best performance in terms of mean squared error (MSE). The model trained with MSE as its loss function and used early stopping to prevent overfitting and trained on 70 percent of the data with the rest being used for test and validation.

The SHAP values are then used to interpret the results of the multi-layer perceptron model. SHAP values are constructed by calculating marginal contribution of a variable as a deviation from the output variable's mean if the variable was changed. This is done across a sample of observations and gives each variable a distribution of SHAP values. The summary plot shows the distribution of SHAP values for the continuous variables.

1. be sure to mention Alpha and how alpha is estimated because it is estimated differently

2. explain some of the strengths and weaknesses of some of the models, particularly lasso

3. explain the neural network part much more in depth

4. explain the reasoning being behind using OLS, as in kitchen sink approach to just see some covariates, but also that it is subject to overfitting

5. explain the advantages and disadvantages of the step wise regression, particularly mention it's advantages and disadvantages in contrast to just standard OLS, along with similarities to OLS

5 Results

The gamma for permanent shocks is a centered heavily around 0 with a standard deviation of 0.1686 as seen in Figure 1. Being centered at 0 is due to its derivation and as seen in Figure 2 there is not a clear correlation with age aside from possibly a slight tightening past the age of 60 as the variance of gamma appears to decrease at least somewhat. To formally quantify these patterns, we next present regression estimates of gamma across several specifications.

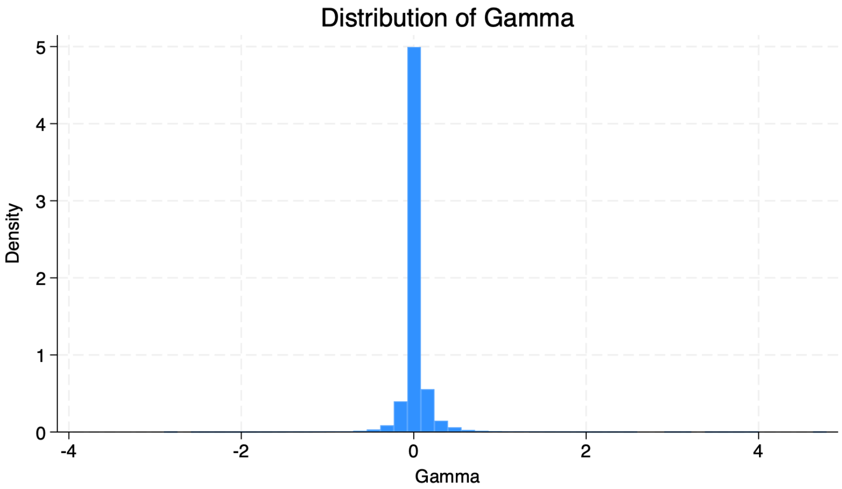

Figure 1: Distribution of Gamma

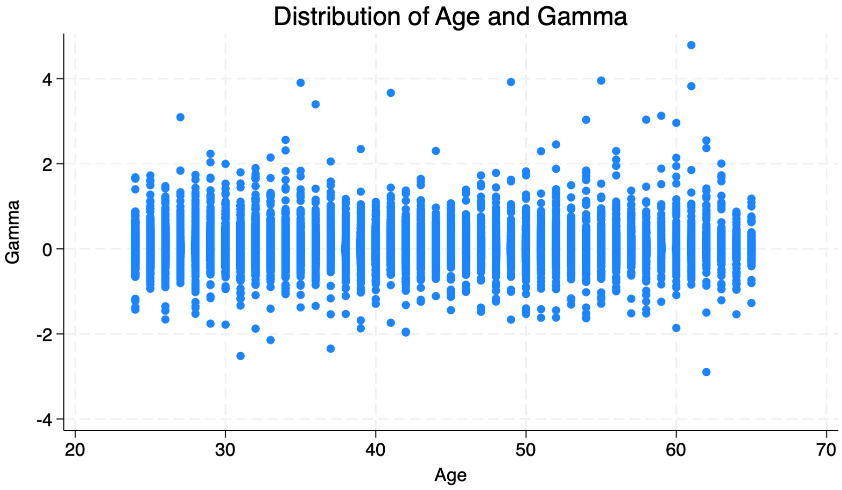

Figure 2: Scatterplot of Age vs. Gamma

Table 1: Gamma Regressions: OLS Results

	(1)	(2)	(3)	(4)	(5)
	Gamma	Gamma	Gamma	Gamma	Gamma
Less than High School	-0.0115***	-0.0107***	-0.0117***	-0.0109***	-0.0117***
	(0.00215)	(0.00232)	(0.00242)	(0.00253)	(0.00255)
High School Graduate	-0.00918***	-0.00854***	-0.00902***	-0.00812***	-0.00850***
	(0.00144)	(0.00149)	(0.00158)	(0.00171)	(0.00173)
Some College	-0.00746***	-0.00681***	-0.00736***	-0.00674***	-0.00695***
	(0.00170)	(0.00172)	(0.00177)	(0.00184)	(0.00185)
Probability of Recession	0.0000422	-0.0104	-0.0000953	-0.0000817	-0.0000292
	(0.0000437)	(0.0179)	(51.52)	(51.50)	(51.49)
Real GDP growth rate	0.000600	-0.00432	-0.0000211	-0.000186	-0.000626
	(0.000400)	(0.00325)	(3.981)	(3.979)	(3.979)
5-year moving average of AEP	0.0000636**	0.0000629*	0.0000374	0.0000766**	0.0000688*
	(0.0000295)	(0.0000322)	(0.0000340)	(0.0000347)	(0.0000356)
Out of Labor Force	-0.00430	-0.00410	-0.00471	-0.00461	-0.00474
	(0.00535)	(0.00536)	(0.00573)	(0.00572)	(0.00572)
Tenure	-0.0000438	-0.0000336	0.0000277	0.00000375	0.0000156
	(0.0000888)	(0.0000963)	(0.0000981)	(0.0000984)	(0.0000987)
Age	0.00592**	0.00650**	0.00679**	0.00679**	0.00671**
	(0.00269)	(0.00273)	(0.00274)	(0.00274)	(0.00274)
Age Squared	-0.000170***	-0.000193***	-0.000196***	-0.000197***	-0.000194***
	(0.0000651)	(0.0000659)	(0.0000662)	(0.0000662)	(0.0000663)
Age Cubed	0.00000153***	0.00000172***	0.00000173***	0.00000173***	0.00000171***
	(0.000000509)	(0.000000515)	(0.000000517)	(0.000000518)	(0.000000518)
State FE	No	Yes	Yes	Yes	Yes
Year FE	No	Yes	Yes	Yes	Yes
Race FE	No	Yes	Yes	Yes	Yes
Cohort FE	No	Yes	Yes	Yes	Yes
Occupation FE	No	No	No	Yes	Yes
Industry FE	No	No	Yes	No	Yes
R-squared	0.001	0.002	0.003	0.005	0.006
N	82357	82333	81556	81556	81556

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The beginning of the analysis is just simple OLS regressions of gamma. The further analysis will focus on which variables are most significant or importance and less on the actual size of the effect. However size of coefficients is something OLS can easily address. While gamma is centered around zero, variables do still have effects despite being small. Table 1 shows the OLS estimates for gamma across different specifications. The different specifications include different sets of controls, such as occupation and industry controls along with other controls such as state, year, cohort, and race.

The coefficients are quite small however some are larger than others. Firstly the edcuation variables are all significant in every model. In addition they also are the largest coefficients outside of the fixed effects in just about every model. The less than high school variable is a clear winner for being the largest averaging around -0.01.

The highschool and some college variables are smaller than less than highschoool which would imply that the common intution is still true and that a high school diploma and college education does provide stabler long term employment and earnings overall.

The other variables that are significant are the age variables. The age, age squared, and age cubed variables are all significant in all models. The interesting result is that the age squared and cubed variables are slightly more significant than the regular age variable. On some level it is expected that towards the end of a person's career they would become more risk averse and it is possible the squared and cubed terms capture this variation occurring right before retirement. Regardless they do imply a nonlinear relation between age and permanent income shocks. Overall the OLS results show that education and age are important in explaining lifetime earnings risk. This pattern continues in the stepwise and lasso results.

Table 2: Gamma Regressions: Stepwise Selection Results

	(1)	(2)	(3)	(4)	(5)
	Gamma	Gamma	Gamma	Gamma	Gamma
Less than High School	-0.0110***	-0.0110***	-0.0101***	-0.0108***	-0.0110***
	(0.00211)	(0.00211)	(0.00205)	(0.00234)	(0.00236)
High School Graduate	-0.00904***	-0.00905***	-0.00840***	-0.00829***	-0.00834***
	(0.00143)	(0.00143)	(0.00143)	(0.00166)	(0.00168)
Some College	-0.00740***	-0.00740***	-0.00731***	-0.00723***	-0.00725***
	(0.00169)	(0.00169)	(0.00172)	(0.00182)	(0.00183)
Age	0.00571*	0.00568*		0.00594*	0.00566*
	(0.00268)	(0.00268)		(0.00270)	(0.00270)
Age Cubed	0.00000150**	0.00000149**	0.000000439***	0.00000150**	0.00000143**
	(0.000000508)	(0.000000508)	(8.71e-08)	(0.000000511)	(0.000000511)
5-year moving average of AEP	0.0000579*	0.0000577*		0.0000788*	0.0000732*
	(0.0000286)	(0.0000286)		(0.0000316)	(0.0000324)
Age Squared	-0.000165*	-0.000165*	-0.0000282***	-0.000168*	-0.000161*
	(0.0000650)	(0.0000650)	(0.00000578)	(0.0000654)	(0.0000654)
Constant	-0.0391	-0.0387	0.0395***	-0.0414	-0.0380
	(0.0358)	(0.0358)	(0.00358)	(0.0360)	(0.0360)
State FE					
Year FE					
Race FE					
Cohort FE					
Occupation FE				✓	✓
Industry FE			✓		✓
State FE Available	No	Yes	Yes	Yes	Yes
Year FE Available	No	Yes	Yes	Yes	Yes
Race FE Available	No	Yes	Yes	Yes	Yes
Cohort FE Available	No	Yes	Yes	Yes	Yes
Occupation FE Available	No	No	No	Yes	Yes
Industry FE Available	No	No	Yes	No	Yes
R-squared	0.001	0.001	0.002	0.003	0.004
N	82357	82333	81556	81556	81556

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3: Gamma Regressions: Lasso Selection Results

	No Controls	No Occ/Ind	No Occ	No Ind	All Controls
EDU1	2	3	3	2	2
EDU2	1	2	2	2	2
EDU3	3	4	4	4	3
OLF	9	9	8	7	6
PrRecess	10				
currentage	8	7	6	5	4
currentagecube	4	8	7	6	5
currentagesq	11	10	10	9	8
ma5aep	5	5	5	3	3
rGDPgrow	6				
tenure	7	6	9	8	7
State FE	No	Yes	Yes	Yes	Yes
Year FE	No	Yes	Yes	Yes	Yes
Race FE	No	Yes	Yes	Yes	Yes
Cohort FE	No	Yes	Yes	Yes	Yes
Occupation FE	No	No	No	Yes	Yes
Industry FE	No	No	Yes	No	Yes

Table 4: Gamma Regressions: Lasso Industry Selection Results

Industry	LASSO Order	SHAP Order
Other Services	1	1
Priv. Househld	1	29
	1	
Legal Services	2	5
Public Administration	3	3
Financial Inst	3	16
Construction	3	23
Health Service	4	6
Clothing/Text	4	14
Mechanical Eng	5	19
Energy/Water	6	11
Agric.,Forestry	7	13
Chemicals	7	15
Educ./Sport	7	17
Retail	8	2
Earth/Clay/Stone	8	20
Social Security	9	21
Volunt./Church	9	24
Wholesale	10	4
Iron/Steel	10	18
Electrical Eng	10	22
Constr. Relate	11	10
Synthetics	11	25
Mining	11	27
Fisheries	12	32
Wood/Paper/Print	13	9
Not Applicable	14	
Other Trans.	15	28
Restaurants	16	12
Train System	16	26
Food Industry	17	7
Insurance	17	31
Service Indust	18	30
Postal System		8

Table 5: Gamma Regressions: Lasso Occupation Selection Results

Occupation	LASSO Order	SHAP Order
Dr./Dentist/Vet	1	11
Insurance Rep.	2	9
Transp. Attend	2	60
	3	
Priv.Bus.Leadr	4	1
Author	4	25
Aero/MarineEngr	4	33
Lumbrman/Axman	5	43
Vendor	6	4
Painter	6	36
Hair Stylist	6	47
Janitor	7	6
Legislator	7	13
Chemist	7	52
Farm Manager	8	10
Music/Perform	8	53
Buyer	8	71
Labor/Craftsmn	9	29
Jewelry Maker	10	63
Mathematician	11	28
Tel. Operator	11	62
Prof. Athlete	11	70
Miner	12	50
Educator	13	5
Machine Fitter	13	34
Stone Cutter	13	72
SecurityServic	14	2
BusinessManagr	14	15
ComputerOperat	14	58
Cook/Waiter	15	32
Fisher/Hunter	15	69
Not Applicable	15	
Agriculturladm	16	45
Eng.Tech.Expert	17	48
ChemicalWorker	17	49
Conductor	17	74
Transport.Oper	18	3
Stenographer	18	66

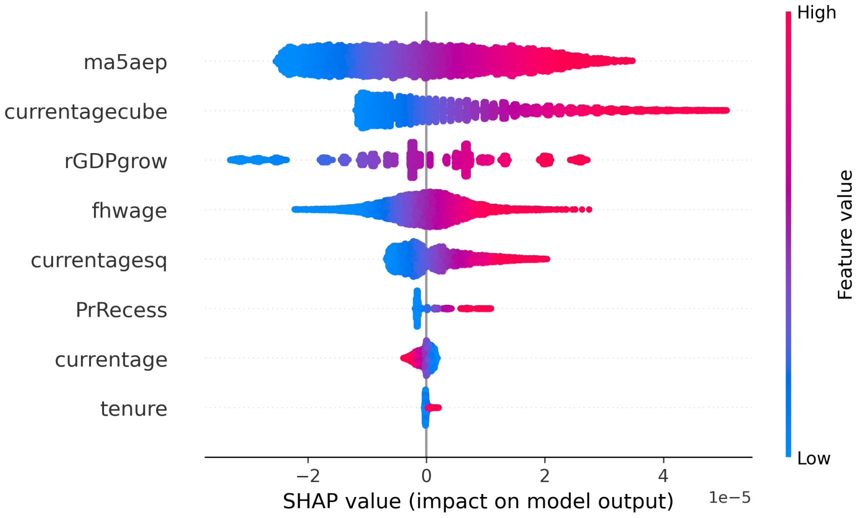

Figure 3: SHAP Summary Plot for Gamma

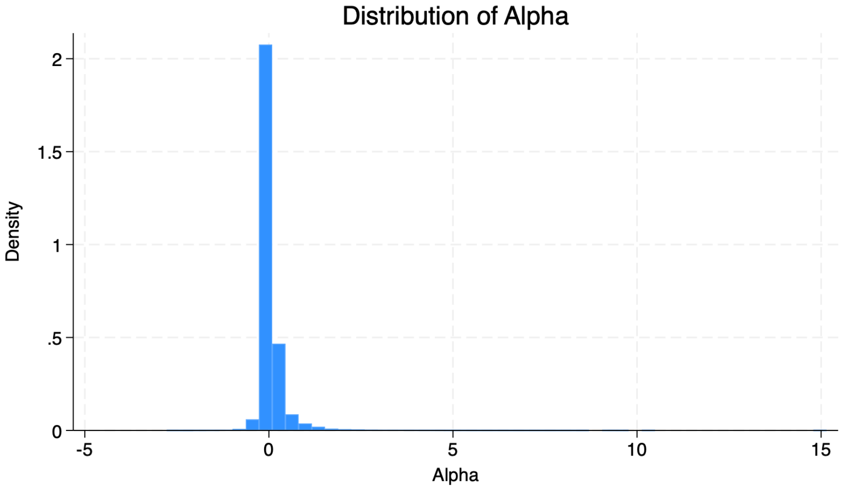

Figure 4: Distribution of Alpha

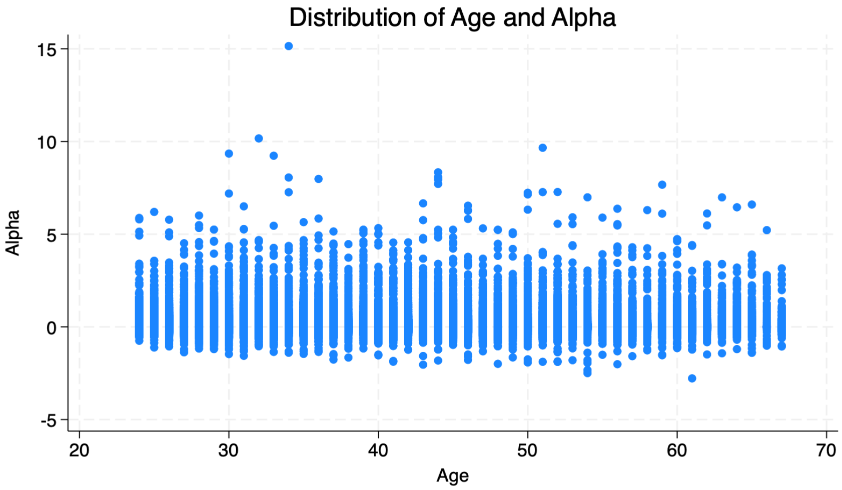

Figure 5: Scatterplot of Age vs. Alpha

Table 6: Alpha Regressions: OLS Results

	(1) Alpha	(2) Alpha	(3) Alpha	(4) Alpha	(5) Alpha
Less than High School	-0.0268***	-0.0240***	-0.0308***	-0.0278***	-0.0301***
	(0.00451)	(0.00487)	(0.00508)	(0.00531)	(0.00535)
High School Graduate	-0.0131***	-0.0129***	-0.0136***	-0.00909**	-0.0104***
	(0.00309)	(0.00320)	(0.00340)	(0.00368)	(0.00372)
Some College	-0.00240	-0.00321	-0.00223	0.00135	0.000562
	(0.00366)	(0.00370)	(0.00384)	(0.00398)	(0.00401)
Probability of Recession	-0.0000420	-0.138**	-0.113	-0.112	-0.113
	(0.0000929)	(0.0554)	(0.0723)	(0.0723)	(0.0722)
Real GDP growth rate	-0.00144*	0.000541	-0.00107	-0.00287	-0.00359
	(0.000855)	(0.00633)	(0.00713)	(0.00710)	(0.00714)
5-year moving average of AEP	0.00127***	0.00138***	0.00103***	0.00111***	0.00106***
	(0.0000616)	(0.0000669)	(0.0000712)	(0.0000726)	(0.0000745)
Out of Labor Force	0.113***	0.111***	0.0518***	0.0521***	0.0517***
	(0.0101)	(0.0101)	(0.0111)	(0.0111)	(0.0111)
Tenure	-0.000698***	-0.00000130	0.000135	0.0000259	0.0000861
	(0.000185)	(0.000202)	(0.000207)	(0.000208)	(0.000208)
Age	0.0165***	0.0183***	0.0175***	0.0182***	0.0173***
	(0.00520)	(0.00526)	(0.00530)	(0.00530)	(0.00530)
Age Squared	-0.000398***	-0.000451***	-0.000431***	-0.000452***	-0.000427***
	(0.000124)	(0.000125)	(0.000126)	(0.000126)	(0.000126)
Age Cubed	0.00000324***	0.00000371***	0.00000356***	0.00000374***	0.00000353***
	(0.000000947)	(0.000000955)	(0.000000963)	(0.000000963)	(0.000000964)
State FE	No	Yes	Yes	Yes	Yes
Year FE	No	Yes	Yes	Yes	Yes
Race FE	No	Yes	Yes	Yes	Yes
Cohort FE	No	Yes	Yes	Yes	Yes
Occupation FE	No	No	No	Yes	Yes
Industry FE	No	No	Yes	No	Yes
R-squared	0.008	0.013	0.020	0.021	0.023
N	102946	102910	100781	100781	100781

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 7: Alpha Regressions: Stepwise Selection Results

	(1)	(2)	(3)	(4)	(5)
	Alpha	Alpha	Alpha	Alpha	Alpha
Less than High School	-0.0261***	-0.0204***	-0.0259***	-0.0259***	-0.0272***
	(0.00418)	(0.00450)	(0.00464)	(0.00477)	(0.00480)
High School Graduate	-0.0122***	-0.0108***	-0.0107***	-0.00852**	-0.00913**
	(0.00271)	(0.00278)	(0.00289)	(0.00304)	(0.00306)
Age Squared	-0.000401**	-0.000394**	-0.000336**	-0.000384**	-0.000351**
	(0.000124)	(0.000124)	(0.000125)	(0.000125)	(0.000125)
Age Cubed	0.00000326***	0.00000327***	0.00000287**	0.00000321***	0.00000296**
	(0.000000947)	(0.000000950)	(0.000000959)	(0.000000959)	(0.000000959)
Real GDP growth rate	-0.00115*				
	(0.000541)				
5-year moving average of AEP	0.00127***	0.00135***	0.00106***	0.00111***	0.00107***
	(0.0000601)	(0.0000650)	(0.0000689)	(0.0000711)	(0.0000728)
Out of Labor Force	0.113***	0.112***	0.0535***	0.0541***	0.0535***
	(0.0101)	(0.0101)	(0.0111)	(0.0111)	(0.0111)
Tenure	-0.000704***				
	(0.000185)				
Age	0.0167**	0.0159**	0.0132*	0.0153**	0.0140**
	(0.00519)	(0.00523)	(0.00527)	(0.00527)	(0.00527)
Constant	-0.190**	-0.157	-0.102	-0.117	-0.105
	(0.0704)	(0.0825)	(0.0831)	(0.0831)	(0.0831)
State FE		✓	✓	✓	✓
Year FE		✓	✓	✓	✓
Race FE		✓	✓	✓	✓
Cohort FE					
Occupation FE				✓	✓
Industry FE			✓		✓
State FE Available	No	Yes	Yes	Yes	Yes
Year FE Available	No	Yes	Yes	Yes	Yes
Race FE Available	No	Yes	Yes	Yes	Yes
Cohort FE Available	No	Yes	Yes	Yes	Yes
Occupation FE Available	No	No	No	Yes	Yes
Industry FE Available	No	No	Yes	No	Yes
R-squared	0.008	0.012	0.018	0.021	0.022
N	102946	102910	100781	100781	100781

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 8: Alpha Regressions: Lasso Selection Results

	No Controls	No Occ/Ind	No Occ	No Ind	All Controls
EDU1	4	5	4	4	4
EDU2	5	6	6	7	7
EDU3	7	7	8	6	6
OLF	2	3	3	3	3
PrRecess	8				
currentage		9	9	8	9
currentagecube	3	4	5	5	5
currentagesq		10	10	9	10
ma5aep	1	2	2	2	2
rGDPgrow	6				
tenure	4	8	7	10	8
State FE	No	Yes	Yes	Yes	Yes
Year FE	No	Yes	Yes	Yes	Yes
Race FE	No	Yes	Yes	Yes	Yes
Cohort FE	No	Yes	Yes	Yes	Yes
Occupation FE	No	No	No	Yes	Yes
Industry FE	No	No	Yes	No	Yes

Table 9: Alpha Regressions: Lasso Industry Selection Results

Industry	LASSO Order	SHAP Order
Not Applicable	1	
Agric.,Forestry	2	2
Fisheries	3	29
	3	
Mechanical Eng	4	1
Public Administration	4	3
Construction	5	13
Legal Services	5	23
Other Services	6	22
Energy/Water	7	15
Constr. Relate	8	31
Clothing/Text	9	16
Synthetics	9	19
Service Indust	10	27
Iron/Steel	11	7
Postal System	12	10
Health Service	13	8
Food Industry	14	17
Chemicals	14	20
Train System	14	24
Educ./Sport	15	9
Retail	16	5
Electrical Eng	16	14
Other Trans.	17	11
Wood/Paper/Print	18	4
Restaurants	18	28
Priv. Househld	18	32
Financial Inst	19	12
Earth/Clay/Stone	20	26
Volunt./Church	21	21
Social Security	22	30
Insurance	23	18
Wholesale		6
Mining		25

Table 10: Alpha Regressions: Lasso Occupation Selection Results

Occupation	LASSO Order	SHAP Order
Not Applicable	1	
Farm Manager	2	1
Fisher/Hunter	3	42
	4	
Mathematician	5	4
Insurance Rep.	6	18
Architect/Engineer	7	8
Inspector	8	9
Bricklay/Carpt	8	41
Agriculturladm	8	44
Priv.Bus.Leadr	9	13
Prof. Athlete	9	55
Music/Perform	10	39
Mailman	11	5
BusinessManagr	11	10
Eng.Tech.Expert	12	15
ComputerOperat	12	40
Hair Stylist	12	74
Ofc.Worker Etc	13	11
Cook/Waiter	13	28
Scientist	14	21
Chemist	15	48
Tech.Salespers	15	59
Economist	15	75
Office Manager	16	32
Aero/MarineEngr	16	38
Sculptr/Paintr	16	58
Printer Etc.	17	25
Rest./StoreMgr	17	52
Jewelry Maker	17	66
Soldier	18	29
Tool/Die Maker	19	6
Administrator	19	36
Stenographer	19	62
ChemicalWorker	20	46
Spinner/Weaver	20	47
Buyer	20	50
Machine Fitter	21	3

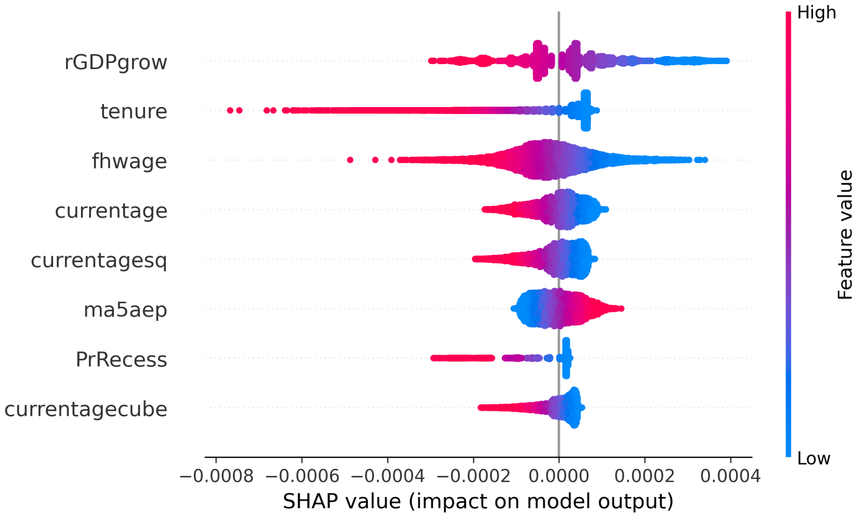

Figure 6: SHAP Summary Plot for Alpha

6 Conclusion

References

[1] Abowd, J. M. and Card, D. (1989). On the Covariance Structure of Earnings and Hours Changes. *Econometrica*, 57(2), 411–445.

[2] Baker, M. (1997). Growth-Rate Heterogeneity and the Covariance Structure of Life-Cycle Earnings. *Journal of Labor Economics*, 15(2), 338–375.

[3] Consumer Price Index for All Urban Consumers (2017). Data set, series CUUR0000SA0. Washington, D.C.: U.S. Bureau of Labor Statistics.

[4] Drewianka, S. (2010). Cross-Sectional Variation in Individuals' Earnings Instability. *Review of Income and Wealth*, 56(2), 291–326.

[5] Drewianka, Scott and Oberg, Phillip. (2025). Earnings Risk and Heterogeneous Expected Earnings Profiles. Quantitative Economics. Forthcoming.

[6] Drewianka, S. and Oberg, P. (2019). Estimating Heterogeneity in Lifetime Earnings Risk. Unpublished manuscript, University of Wisconsin–Milwaukee and Illinois Wesleyan University. Available at `http://dx.doi.org/10.2139/ssrn.3630477`.

[7] Guvenen, F. (2009). An Empirical Investigation of Labor Income Processes. *Review of Economic Dynamics*, 12, 58–79.

[8] Guvenen, F. (2007). Learning Your Earning: Are Labor Income Shocks Really Very Persistent? *American Economic Review*, 97(3), 687–712.

[9] Hill, M. S. (1992). *The Panel Study of Income Dynamics: A User's Guide*, Volume 2. Newbury Park, CA: Sage Publications.

[10] Hryshko, D. (2012). Labor Income Profiles Are Not Heterogeneous: Evidence from Income Growth Rates. *Quantitative Economics*, 3, 177–209.

[11] Lundberg, S. M., and Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

[12] MaCurdy, T. E. (1982). The Use of Time Series Processes to Model the Error Structure of Earnings in a Longitudinal Data Analysis. *Journal of Econometrics*, 18(1), 83–114.

[13] Meghir, C. and Pistaferri, L. (2004). Income Variance Dynamics and Heterogeneity. *Econometrica*, 72(1), 1–32.

[14] Panel Study of Income Dynamics (2023). Public use dataset. Ann Arbor, MI: Institute for Social Research, Survey Research Center, University of Michigan.